frontiers

Machine Learning techniques for the diagnosis of Schizophrenia based on Event Related Potentials

Elsa Santos Febles[1*], Marlis Ontivero Ortega[1], Michell Valdés Sosa[1], Hichem Sahli[2,3]

[1]Cuban Neuroscience Center, Havana, Cuba
[2]Department of Electronics and Informatics (ETRO), Vrije Universiteit Brussel (VUB), Brussels, Belgium
[3]Interuniversity Microelectronics Centre (IMEC), 3001 Leuven, Belgium

*** Correspondence:**
Elsa Santos Febles
elsa@cneuro.edu.cu

Keywords: Multiple Kernel Learning, schizophrenia, Boruta, feature selection, event related potential, machine learning

Abstract

Antecedent: The diagnosis of schizophrenia could be enhanced with objective neurophysiological biomarkers, such as the event related potential features in conjunction with machine learning procedures. A previous work extracted features from event related responses to three oddball paradigms (auditory and visual P300, and mismatch negativity) for the discrimination of schizophrenic patients. They used several classifiers: Naïve Bayes, Support Vector Machine, Decision Tree, Adaboost and Random Forest. The best accuracy was obtained with Random Forest (84.7%).

Objective: The aim of this study was to examine the efficacy of Multiple Kernel Learning classifiers and Boruta feature selection method exploring different features for single-subject classification between schizophrenia patients and healthy controls.

Methods: A cohort of 54 schizophrenic subjects and 54 healthy control subjects were studied. Three sets of features related to the event related potentials signal were calculated: Peak related features, Peak to Peak related features and Signal related features. The Boruta feature selection algorithm was used to evaluate its impact on classification accuracy. A Multiple Kernel Learning algorithm was applied to address schizophrenia detection.

Results: We obtained a classification accuracy of 83% using Multiple Kernel Learning classifier with the whole dataset. This result in accuracy triangulates previous work and shows that the differences between schizophrenic patients and controls are robust even when different classifiers are used. Appling the Boruta feature selection algorithm a classification accuracy of 86% was yielded. The variables that contributed most to the classification were mainly related to the latency and amplitude of the auditory P300.

Conclusion: This study showed that Multiple Kernel Learning can be useful in distinguishing between schizophrenic patients and controls. Moreover, the combination with the Boruta algorithm provides an improvement in classification accuracy and computational cost.

1 Introduction

Schizophrenia is a severe and persistent debilitating psychiatric disorder with prevalence of 1% of the world population (McGrath et al., 2004). Although psychotic symptoms such as hallucinations and delusions are frequently present, impaired information processing is probably the most common symptom (Javitt et al., 1993). This deficit is reflected mainly in deficits in attention and working memory tasks when compared with healthy controls (Li et al., 2018). The diagnosis of schizophrenia is made by psychiatrists by ascertaining the presence of predefined symptoms (or their precursors) with personal interviews. However, in some cases this diagnosis is unclear, or patients are misdiagnosed with Schizophrenia (Coulter et al., 2019). Thus, finding biomarkers for the prediction of individuals with schizophrenia would be desirable in order to choose the optimal treatment (pharmacologic or non-pharmacologic). Analysis of EEG recording during information processing tasks could provide objective complimentary measures to support the subjective human-based decision process (Sabeti et al., 2009; Koukkou et al., 2018) .

EEG is a non-invasive and low-cost technique used to measure electrical brain activity along multiple scalp locations. EEG signals have been widely adopted to study mental disorders, such as dementia, epileptic seizures, cognitive dysfunction, among others, as well as schizophrenia (Loo et al., 2016; Olbrich et al., 2016; Horvath et al., 2018). EEG reflects the spontaneous activity of myriad brain parcels, but also can include responses to afferent stimuli (Cong et al., 2015). Event related potentials (ERPs) are electrical responses that are time-locked to a specific stimulus or event, and can be used to assess brain dynamics during information processing in specific tasks (Woodman, 2010). When a subject is presented with a series of standard stimuli, interspersed with infrequent deviant stimuli, the Mismatch Negativity (MMN) (Lee et al., 2017) and the P300 (Li et al., 2018) components are generated. This task is known as the oddball paradigm and is used to study schizophrenia since consistent deficits in the P300 and MNN have been reported in this disease (Bramon et al., 2004; Javitt et al., 2017). Although MMN and P300 are usually produced by an infrequent unexpected event in a sequence of auditory stimuli, P300 can also be obtained with visual stimuli. The MMN is of shorter latency and does not require attention to the stimulus (Näätänen et al., 2004), whereas the P300 is of longer latency and requires attention to the stimulus (Huang et al., 2015).

Several studied have reported significant differences in the latency and amplitude of MMN and P300 between controls and patients, suggesting that these features are possible markers of the prodromal phase of schizophrenia (Atkinson et al., 2012; Loo et al., 2016) as well as a potential endophenotypes for schizophrenia (Earls et al., 2016). Analysis of a large dataset of auditory P300 ERP (649 controls and 587 patients) confirmed the reliability of this reduced amplitude, with a large effect size (Turetsky et al., 2015). However, these findings of statistically significance differences in a group analysis does not imply that EEG is useful for the prediction of individual schizophrenia cases (Lo et al., 2015), which requires applying a prediction paradigm using Machine Learning.

Accordingly, machine learning techniques are being applied to classify between schizophrenics (SZs) and healthy controls (HCs) using ERPs. The most common features used are based on amplitude and latency of different components (e. g. N100 and P300 (Neuhaus et al., 2013), P50 and N100 (Iyer et al., 2012; Neuhaus et al., 2014)), with several classifiers tested. Neuhaus et al. using visual and

auditory oddball paradigms and a k-nearest neighbor (KNN) classifier obtained a classification accuracy of 72.4 % (Neuhaus et al., 2013). The same author with a bigger sample size and a Naive Bayes (NB) classifier achieved a 77.7% of accuracy (Iyer et al., 2012). Laton et al. evaluated the performance of several classifiers extracting features from auditory/visual P300 and MMN (Laton et al., 2014). The results using NB and Decision Tree (without and with AdaBoost) achieved accuracies of about 80%. Recently, Barros et al. published a critical review that summarizes machine learning-based classification studies to detect SZs based on EEG signals, conducted since 2016, (Barros et al., 2021). These authors reported that Support Vector Machines (SVM) were the commonly used algorithms, probably due to its computational efficiency. This kernel-based learning method also achieved the best performance in most studies. Nevertheless, none of the studies focused on ERPs, have used multiple kernels, employing instead only one specific kernel function.

The multiple kernel learning (MKL) method learns a weighted combination of different kernel functions and is able to benefit from information coming from multiple sources (Wani and Raza, 2018). It has been used to address the problem of biomarker evaluation for schizophrenia detection, but basically applied to Magnetic Resonance Images increasing performance accuracy (Ulaş et al., 2012; Castro et al., 2014; Liu et al., 2017). However, as far as we know, application of MKL to electrophysiological data has been not explored for schizophrenia, even though some authors are applying this technique to EEG signals for other purposes, mainly brain computer interfaces (Li et al., 2014; Zhang et al., 2017). Thus, MKL has not been applied in the objective diagnosis of Schizophrenia using EEG.

Here, using the same dataset provided by Laton et al. (Laton et al., 2014), we extended the set of predictor variables beyond the latency and amplitude of the ERP components, by including additional morphological features (based on time) together with some features extracted from the frequency domain. Due to the large number of features, the Boruta method was applied, which is a wrapper Random Forest (RF) based feature selection algorithm, to estimate the impact of a subset of important and relevant feature variables in the classification accuracy. The multiple kernel learning (MKL) was evaluated for the classification of SZs versus HCs.

2 Materials and methods

2.1 Dataset

The study was carried out on data from 54 patients and 54 controls, matched for age and gender. Patients were classified by a semi-structured interview (OPCRIT v4.0) and all participants gave written informed consent. Detailed demographic data can be found in **Table 1**. EEGs were recorded using a 64-channel and the international 10/10 system, with a sampling frequency of 256 Hz. Three paradigms auditory/visual P300 and MMN were used. **Table 2** shows a brief description of paradigms.

The signals were filtered using bandpass Butterworth filters with cuttoffs at 0.1 and 30 Hz. Epochs were extracted using time windows between -200 and 800 ms for the P300 paradigms, and between -100 and 500 ms for the MMN. Subsequently, baseline correction, re-referencing to linked ears and artefact rejection were performed. Finally, epochs were averaged into stimulus specific responses for each individual and low-pass filter and baseline correction were re-applied. More details can be found in Laton et al (Laton et al., 2014).

2.2 Feature extraction

Feature extraction has been carried out on the waveform of ERPs emerged as the averaging of the electrical responses corresponding to the set of stimuli different from the standard stimulus (Target and Distractor for P300, Duration and Deviant for MNN)). Only Fz, Cz and Pz channels were considered (see **Figure 1**). Thus, the number of features extracted for classification purposes was 726 (282 features for each P300 paradigms and 162 for MMN paradigm). The feature values were standardized to ensure that all of them have equal weight during training of the classifiers. These standardized values were then normalized, rescaling them all to values between 0 and 1. In this binary classification problem, patients and controls were 1 and 0 respectively.
The set of features can be divided into three categories: Peak related features, Peak to Peak related features and Signal related features. Details about feature definitions are presented in **Annex 1**. Some of these features were previously used for other authors to calculate features related to the ERP signal (Kalatzis et al., 2004; Abootalebi et al., 2009). Four peaks for P300 paradigms (N100, P200, N200, and P300) and two peaks for MMN paradigm (N200, P300) were considered (see **Figure 2**).

2.2.1 Peak related features

Peaks were estimated using the same algorithm described in Laton et al (Laton et al., 2014). The algorithm detects the largest absolute value in an interval established around the average latency of the peak in the respective grand average. This value is considered as *Amplitude* of the corresponding peak, their *Latency* is the time where the peak appears in the respective time interval. To ensure little overlap between the intervals, the detection interval was extended to contain the latency of peak most deviated. To search the latency of the peak, the minimum value of the corresponding detection interval was changed by the latency of the previous searched peak to avoid mistakes in the order of the ERPs components. The other features were: *Absolute Amplitude, Latency/Amplitude ratio, Absolute Latency/Amplitude ratio, Average Absolute Signal Slope* and *Slope sign alterations*.

2.2.2 Peak to Peak related features

Three features were calculated considering the relationship between adjacent selected peaks: the absolute difference between the amplitude of the peak and the next peak in latency order; the difference in latencies of these two peaks; and the slope of the signal in this time window.

2.2.3 Signal related features

Features considering the area under the curve were calculated: the sum of the positive signal values (*Positive Area*); the sum of the negative signal values (*Negative Area*); the *Total Area*, and *Absolute Total Area*. Two more features related to the whole signal were calculated: the number of times that the amplitude value of the signal crosses the zero y-axis between two adjacent peaks (*Zero Crossing*); and the relation of the number of crosses per time interval (*Zero Cross Density*).

Additionally, frequency domain features were extracted using a Power Spectral Density (PSD) analysis: the frequency with the largest energy content in the signal (*Mode frequency*) spectrum; the frequency that separates the power spectrum into two equal energy areas (*Median frequency*); and an estimate of the central tendency of the derivate power distributions (*Mean frequency*).

2.3 Classifier used in the study

2.3.1 MKL

The use of MKL has shown that it enhances the interpretability of decision functions and can improve classification performance compared with other classifiers (Kloft et al., 2009; Varma and Babu, 2009). Similar to simple SVM applications, this method is based on kernel definitions, however, instead of one single kernel, MKL combines several kernel functions (reflecting different

kinds of information), and also automatically determines the importance of each kernel (Gönen and Alpaydin, 2011) .

Given a set of data X and a feature mapping function Φ, a kernel matrix can be defined as the inner product of each pair of feature vectors:

$$\boldsymbol{K}(\boldsymbol{x_i}, \boldsymbol{x_j}) = \langle \Phi(\boldsymbol{x_i}), \Phi(\boldsymbol{x_j}) \rangle$$

In the multiple kernel learning problem for binary classification, $\boldsymbol{N}$ data points $(\boldsymbol{x_i}, \boldsymbol{y_i})$ $(\boldsymbol{y_i} \epsilon \pm \boldsymbol{1})$ are given, where $\boldsymbol{x_i}$ is translated via $\boldsymbol{M}$ mappings $\Phi_{\boldsymbol{m}}(\boldsymbol{x}) \rightarrow \mathbf{R}^{\boldsymbol{D_m}}$, $\boldsymbol{m= 1,\ldots,M}$, from the input into $\boldsymbol{M}$ feature spaces $\Phi_{\boldsymbol{1}}(\boldsymbol{x_i}), \ldots., \Phi_{\boldsymbol{M}}(\boldsymbol{x_i})$ where $\boldsymbol{D_m}$ denotes the dimensionality of the $\boldsymbol{m^{th}}$ feature space.

Multiple Kernel Learning methods aim to construct an optimal kernel model where the kernel is a linear combination of fixed base kernels. Learning the kernel then consists of learning the weighting coefficients **β** for each base kernel, rather than optimizing the kernel parameters of a single kernel.

$$\boldsymbol{K_{opt}}(\boldsymbol{x_i}, \boldsymbol{x_j}) = \sum_{\boldsymbol{m=0}}^{\boldsymbol{M}} \boldsymbol{\beta_m} \boldsymbol{K_m}(\boldsymbol{x_i}, \boldsymbol{x_j}) \qquad \boldsymbol{\beta_k > 0}, \sum_{\boldsymbol{m=0}}^{\boldsymbol{M}} \boldsymbol{\beta_m} = \boldsymbol{1}$$

When MKL is plugged into SVM, the primal form of MKL is reformulated as the following optimization problem:

$$\min_{\boldsymbol{\beta}, \boldsymbol{w}, \boldsymbol{b}, \boldsymbol{\mathcal{E}}} \frac{\boldsymbol{1}}{\boldsymbol{2}} \sum\nolimits_{\boldsymbol{m}} \frac{\boldsymbol{1}}{\boldsymbol{\beta_m}} \|(\boldsymbol{w_m})\|^{\boldsymbol{2}}_{\boldsymbol{\mathcal{H}_m}} + \boldsymbol{C} \sum\nolimits_{\boldsymbol{i}} \mathcal{E}_{\boldsymbol{i}}$$

$$\boldsymbol{s.t} \qquad \boldsymbol{y_i} \left(\sum\nolimits_{\boldsymbol{m}} \langle (\boldsymbol{w_m}), \boldsymbol{\phi_m}(\boldsymbol{x_i}) \rangle_{\boldsymbol{\mathcal{H}_m}} + \boldsymbol{b} \right) + \mathcal{E}_{\boldsymbol{i}} \geq \boldsymbol{1}$$

$$\mathcal{E}_{\boldsymbol{i}} \geq \boldsymbol{0}, \boldsymbol{for} \; \forall \boldsymbol{i}$$

$$\sum\nolimits_{\boldsymbol{m}} \boldsymbol{\beta_m} = \boldsymbol{1}, \qquad \boldsymbol{\beta} \geq \boldsymbol{0}$$

where $\boldsymbol{C}$ is a regularization parameter between training errors and an optimal separating hyperplane. For binary classification MKL problem, optimization is solved using semi-infinite programming (Sonnenburg et al., 2006). The three commonly used kernels are: linear kernel ($\boldsymbol{K_L}$), polynomial kernel ($\boldsymbol{K_P}$), and Gaussian kernel ($\boldsymbol{K_g}$):

$$\boldsymbol{K_L}(\boldsymbol{x_i}, \boldsymbol{x_j}) = \langle \boldsymbol{x_i}, \boldsymbol{x_j} \rangle$$

$$\boldsymbol{K_P}(\boldsymbol{x_i}, \boldsymbol{x_j}) = (\langle \boldsymbol{x_i}, \boldsymbol{x_j} \rangle + \boldsymbol{1})^{\boldsymbol{q}}$$

$$\boldsymbol{K_G}(\boldsymbol{x_i}, \boldsymbol{x_j}) = \boldsymbol{\exp}\left(- \frac{\|\boldsymbol{x_i} - \boldsymbol{x_j}\|^{\boldsymbol{2}}}{\boldsymbol{s^2}} \right),$$

where parameter $\boldsymbol{q}$ is the polynomial degree and parameter $\boldsymbol{s}$ determines the width for Gaussian distribution.

MKL provides a general framework for learning from multiple and heterogeneous data sources (de Carvalho, 2019). This machine learning algorithm works by first constructing a kernel from each of the data sources and then combining these kernels based on a certain criterion for improved classification performance. With ***M*** kernels, a given input data can be mapped into ***M*** feature spaces. Another approach is when different basis kernels are applied to the same data features to identify which kernel is best for the problem at hand.

In this paper, the input data was mapped into different feature spaces trying to group variables with common aspects: type of paradigm, channels (Fz, Cz, Pz), or type of feature. For every feature space the 726 features were rearranged in three groups considering the common aspects (see **Figure 3**). Then, the MKL available in SHOGUN toolbox was applied (Sonnenburg et al., 2010) for every feature space. We used a non-sparse MKL with L2-norm that have more advantages over sparse integration method for thoroughly combining complementary information in heterogeneous data sources. L2-norm distributes the weights over all kernels while taking advantages of the effects of kernels in the objective optimization (Yu et al., 2010).

2.4 Feature Selection

Feature selection yields a subset of features from the original set of features, which are the best representatives of the data. Therefore, it allows us to reduce the number of input variables. The goal of this process is to reduce the computational cost when developing a predictive model and, in some cases, to improve the performance of the model, not always guarantee (Benouini et al., 2020).

2.5 Boruta algorithm

Boruta is a feature selection algorithm that uses a wrapper method based on the RF classifier to measure the importance of variables. RF makes it relatively fast due to its simple heuristic feature selection procedure (Kursa, 2017). In the Boruta algorithm, the original feature set is extended by adding shadow variables (Kursa and Rudnicki, 2010). A shadow variable is created by shuffling values of the original feature. The importance values are calculated for all the attributes by running RF classifier resulting in a *Z score*. The maximum *Z score* is calculated among those shadow variables to assign a hit for each feature that scored better than this maximum. A two-sided test of equality is performed to obtain a statistically significant division between relevant and unimportant feature variables. If a variable systematically falls below the shadow ones, its contribution to the model is doubtful and is therefore eliminated. The shadow variables are removes and the process continues until all variables are accepted, rejected or a limit number of iterations is reached. This limit corresponds to the maximal number of RF runs.

The package “Boruta” in R was used (Kursa and Rudnicki, 2020). The implementation defaults to 100 as the maximum number of RF runs. To get a reduced number of attributes left undecided, this value was set to 500. Nevertheless, when this value isn’t enough, another function *TentativeRoughFix*, contained in the package, can be used to analyses those attributes which importance is very close to the decision criteria.

2.6 Nested cross validation

For explore the feature selection impact, nested cross validation (NCV) was applied. The NCV is characterized by having an inner loop responsible for model selection/hyperparameter tuning and an outer loop is for error estimation. The entire data was divided randomly into ***k*** subsets or folds with stratification, the same proportion of patients and controls as in the complete dataset. The ***k-1*** subsets

are used for feature selection and the remaining subset for testing the model after feature selection. As in k-fold cross-validation method, this process was repeated ***k*** times (outer loop), each time leaving out one of the subsets reserved for testing and the rest for feature selection using Boruta algorithm (see **Figure 4**).

Each subset obtained after feature selection, was used for model hyperparameter tuning in the inner loop. One of the approaches commonly used in practice for the selection of hyperparameters is to try several combinations of them and evaluate their out of sample performance. The tuned parameters in the MKL classifier were:

- Regularized parameter ***C***, a tradeoff between misclassification and simplicity of the model, the candidate's values for grid was 0.5, 1, 1.5 ,5, 10
- Type of kernel (linear, RBF, and polynomial)
- In case of RBF kernels the Sigma (σ) to determine the width for Gaussian distribution, exploring the following values 10, 5, 1, 0.25, 0.5, 0.75.

The parameter configuration selected to train the final model was the one that reached the highest average accuracy on the inner loop. The whole dataset used for tuning parameters was then trained and tested with its corresponding test set in the outer loop. The classifiers' performance was obtained by averaging the accuracy of the ***k*** trained models.

3 Results

3.1 Feature Selection

The Boruta algorithms yielded an average of 32 attributes selected per ***k*** iteration with values in a range of 26 to 42 (see **Figure 5A**). The median computation times was around 2.6 minutes (std 0.04), with 0.005 min per RF runs. A total of 76 attributes were selected at least once. **Figure 5B** shows how many times these attributes were selected in the process. The distribution of variable per paradigm is also showed. The 80% of the 76 attributes selected were related to amplitude, latency, or the correlation between them. Attributes related to frequency domain was barely selected.

Only seven features were identified as important every time Boruta algorithm was used. **Table 3** describes these features according to the paradigm, type of stimulus, channel, and type of feature.

3.2 Classifier performance

To compare the performance of the MKL algorithms three metrics derived from the confusion matrix were used. As the classes were balanced, accuracy (*Acc*) is a good measure for assessing the classification models. Accuracy is the proportion of the total number of predictions that were correct. The other two measures were sensibility (*Sen*) that evaluates true positive rates, and specificity (*Spe*) to evaluate the false positives rates. In **Table 4**, the performance of MKL algorithm when feature selection was applied or not is shown.

3.3 Discussion

Here we explored the use of MKL classification algorithm for distinguishing SZs from HCs based on ERP data. Using all features, the best classification accuracy (83%) was achieved when kernels were built by grouping features according to paradigms. Moreover, when MKL was combined with Boruta method, a classification accuracy of 86% was obtained. With this feature selection algorithm, the

large number of predictor variables was reduced significantly (96%) with a lower computation time. Therefore, training time of MKL was also reduced, its main shortcoming is known to be its high computational cost, especially when many features are used (de Carvalho, 2019).

Review of the Boruta algorithm results pointed out that variables with major importance were mainly related with auditory P300 ERP paradigm. This correspond with the general finding that the P300 measures obtaining from auditory stimuli are more effective in differentiating SZs from HCs than those obtaining from visual stimuli (Park et al., 2005). An interesting point to be noted is that feature selected by Boruta were mainly related with amplitude, latency, and correlation between them. These features correspond with Peak related features. To a lesser extend Peak to Peak related features was included in the selection. However, only three features of Signal related features were rarely included, thus features in frequency domain didn´t contribute to classification.

Overall, these findings are in accordance with findings reported by other authors, and thus triangulates the previous results and shows that the differences between SZs and HCs are robust even when different classifiers are used. Numerous authors have been concluded that odd-ball tasks are potential biomarker for diagnosis in schizophrenia. Some of them have verified that the use of latency and amplitude produces similar results in the discrimination of SZs from HCs. Santos-Mayo et al. used time and frequency ERP features, they explored several electrodes grouping, classifiers, feature selection algorithms and filtering schemes (Santos-Mayo et al., 2017). They achieved accuracies above 90% but their dataset was unbalanced and small, which could limit the generalization of their findings. Shim et al. proposed to extend P300 amplitude and latency sensor-level feature with cortical current density values as source-level features, due to the low spatial resolution originating from volume conduction (Shim et al., 2016). Using Fisher's scores, feature set ranged for 1 to 20 were selected for classification. They reported classification accuracies of 81% for sensor-level features, 85% for source-level features and 88% combined them, using SVM classifier. Laton et al. combined latency and amplitude features of responses to three different odd-ball tasks to apply several classification algorithms (Laton et al., 2014). They achieved a classification accuracy averaged 77% (3.5 std) and their best result, closed to 85%, corresponded to RF classifier. These authors also found a similar pattern in terms of the most relevant features, since in a ranking of the 20 main variables, 14 were extracted from the P300 auditory oddball paradigm. They stated auditory P300 as the most valuable of the three ERP paradigms to the final prediction success.

Compared with these previous studies, our accuracies values are in a range considered as a good accuracy, very close to the results previously reached. This result adds robustness to the previous findings remarking the possibility of accurately distinguish SZs from HCs using neurophysiological measurements. The present finding confirms that Boruta algorithm is a computationally efficient and robust algorithm that improves classification accuracy in many scenarios (Speiser et al., 2019).

Although the approach used here meet our goals, the information of the spatial voltage distributions over the scalp surface was wasted. It is known that the topography across the scalp was significantly different between schizophrenia and normal control groups (Morstyn et al., 1983; Frantseva et al., 2014). Some authors had investigated the topographic abnormalities of schizophrenia mainly group-based researches (Basile et al., 2004). However, individual patient-level analysis using topographic features has been less explored for schizophrenia. This would be a fruitful area for further work in other to reliably classify SZs from HCs.

This study suffers of small sample size as usual in psychiatric cohorts. In these cases, instead of a-priori train/validate/test partitions, strategies of cross-validation allow to estimate the selected model

313 performance and avoid the risk of data leakage. Nevertheless, larger sets yield a more stable, reliable
314 estimate of future performance and guarantee better generalization (Cearns et al., 2019).

315 **4 References**

316 Abootalebi, V., Moradi, M. H., and Khalilzadeh, M. A. (2009). A new approach for EEG feature
317 extraction in P300-based lie detection. *Comput. Methods Programs Biomed.* 94, 48–57.
318 doi:10.1016/j.cmpb.2008.10.001.

319 Atkinson, R. J., Michie, P. T., and Schall, U. (2012). Duration mismatch negativity and P3a in first-
320 episode psychosis and individuals at ultra-high risk of psychosis. *Biol. Psychiatry* 71, 98–104.
321 doi:10.1016/j.biopsych.2011.08.023.

322 Barros, C., Silva, C. A., and Pinheiro, A. P. (2021). Advanced EEG-based learning approaches to
323 predict schizophrenia: Promises and pitfalls. *Artif. Intell. Med.* 114.
324 doi:10.1016/j.artmed.2021.102039.

325 Basile, L. F. H., Yacubian, J., Ferreira, B. L. C., Valim, A. C., and Gattaz, W. F. (2004). Topographic
326 abnormality of slow cortical potentials in Schizophrenia. *Brazilian J. Med. Biol. Res.* 37, 97–
327 109. doi:10.1590/S0100-879X2004000100014.

328 Benouini, R., Batioua, I., Ezghari, S., Zenkouar, K., and Zahi, A. (2020). Fast feature selection
329 algorithm for neighborhood rough set model based on Bucket and Trie structures. *Granul.*
330 *Comput.* 5, 329–347. doi:10.1007/s41066-019-00162-w.

331 Bramon, E., Rabe-Hesketh, S., Sham, P., Murray, R. M., and Frangou, S. (2004). Meta-analysis of
332 the P300 and P50 waveforms in schizophrenia. *Schizophr. Res.* 70, 315–329.
333 doi:10.1016/j.schres.2004.01.004.

334 Castro, E., Gómez-verdejo, V., Martínez-ramón, M., Kiehl, K. A., and Calhoun, V. D. (2014).
335 NeuroImage A multiple kernel learning approach to perform classi fi cation of groups from
336 complex-valued fMRI data analysis : Application to schizophrenia. *Neuroimage* 87, 1–17.
337 doi:10.1016/j.neuroimage.2013.10.065.

338 Cearns, M., Hahn, T., and Baune, B. T. (2019). Recommendations and future directions for
339 supervised machine learning in psychiatry. *Transl. Psychiatry*. doi:10.1038/s41398-019-0607-2.

340 Cong, F., Ristaniemi, T., and Lyytinen, H. (2015). *Advanced signal processing on brain event-*
341 *related potentials: Filtering ERPs in time, frequency and space domains sequentially and*
342 *simultaneously*. doi:10.1142/9306.

343 Coulter, C., Baker, K. K., and Margolis, R. L. (2019). Specialized consultation for suspected recent-
344 onset schizophrenia: Diagnostic clarity and the distorting impact of anxiety and reported
345 auditory hallucinations. *J. Psychiatr. Pract.* 25, 76–81. doi:10.1097/PRA.0000000000000363.

346 de Carvalho, J. A. A. L. (2019). Is Multiple Kernel Learning better than other classifier methods ?

347 Earls, H. A., Curran, T., and Mittal, V. (2016). A Meta-analytic Review of Auditory Event-Related
348 Potential Components as Endophenotypes for Schizophrenia: Perspectives from First-Degree
349 Relatives. *Schizophr. Bull.* 42, 1504–1516. doi:10.1093/schbul/sbw047.

Frantseva, M., Cui, J., Farzan, F., Chinta, L. V., Perez Velazquez, J. L., and Daskalakis, Z. J. (2014). Disrupted cortical conductivity in schizophrenia: TMS-EEG study. *Cereb. Cortex* 24, 211–221. doi:10.1093/cercor/bhs304.

Gönen, M., and Alpaydin, E. (2011). Multiple kernel learning algorithms. *J. Mach. Learn. Res.* 12, 2211–2268.

Horvath, A., Szucs, A., Csukly, G., Sakovics, A., Stefanics, G., and Kamondi, A. (2018). EEG and ERP biomarkers of Alzheimer's disease: A critical review. *Front. Biosci. - Landmark* 23, 183–220. doi:10.2741/4587.

Huang, W. J., Chen, W. W., and Zhang, X. (2015). The neurophysiology of P 300 - An integrated review. *Eur. Rev. Med. Pharmacol. Sci.* 19, 1480–1488.

Iyer, D., Boutros, N. N., and Zouridakis, G. (2012). Clinical Neurophysiology Single-trial analysis of auditory evoked potentials improves separation of normal and schizophrenia subjects. *Clin. Neurophysiol.* 123, 1810–1820. doi:10.1016/j.clinph.2011.12.021.

Javitt, D. C., Doneshka, P., Zylberman, I., Ritter, W., and Vaughan, H. G. (1993). Impairment of early cortical processing in schizophrenia: An event-related potential confirmation study. *Biol. Psychiatry* 33, 513–519. doi:10.1016/0006-3223(93)90005-X.

Javitt, D. C., Lee, M., Kantrowitz, J. T., and Martinez, A. (2017). Mismatch negativity as a biomarker of theta band oscillatory dysfunction in schizophrenia. *Schizophr. Res.* doi:10.1016/j.schres.2017.06.023.

Kalatzis, I., Piliouras, N., Ventouras, E., Papageorgiou, C. C., Rabavilas, A. D., and Cavouras, D. (2004). Design and implementation of an SVM-based computer classification system for discriminating depressive patients from healthy controls using the P600 component of ERP signals. *Comput. Methods Programs Biomed.* 75, 11–22. doi:10.1016/j.cmpb.2003.09.003.

Kloft, M., Brefeld, U., Laskov, P., Müller, K.-R., Zien, A., and Sonnenburg, S. (2009). Efficient and Accurate Lp-Norm Multiple Kernel Learning. *Adv. Neural Inf. Process. Syst.*, 997–1005.

Koukkou, M., Koenig, T., Banninger, A., Rieger, K., Hernández, L. D., Higuchi, Y., et al. (2018). "Neurobiology of Schizophrenia: Electrophysiological Indices," in *Advances in Psychiatry* (Springer, Cham), 433–459. doi:10.1007/978-3-319-70554-5.

Kursa, M. B. (2017). Efficient all relevant feature selection with random ferns. *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)* 10352 LNAI, 302–311. doi:10.1007/978-3-319-60438-1_30.

Kursa, M. B., and Rudnicki, W. R. (2010). Feature selection with the boruta package. *J. Stat. Softw.* 36, 1–13. doi:10.18637/jss.v036.i11.

Kursa, M. B., and Rudnicki, W. R. (2020). Package 'Boruta.' 1–17.

Laton, J., Van Schependom, J., Gielen, J., Decoster, J., Moons, T., De Keyser, J., et al. (2014). Single-subject classification of schizophrenia patients based on a combination of oddball and mismatch evoked potential paradigms. *J. Neurol. Sci.* 347, 262–267.

doi:10.1016/j.jns.2014.10.015.

Lee, M., Sehatpour, P., Hoptman, M. J., Lakatos, P., Dias, E. C., Kantrowitz, J. T., et al. (2017). Neural mechanisms of mismatch negativity dysfunction in schizophrenia. *Mol. Psychiatry* 22, 1585–1593. doi:10.1038/mp.2017.3.

Li, F., Wang, J., Jiang, Y., Si, Y., Peng, W., Song, L., et al. (2018). Top-down disconnectivity in schizophrenia during P300 tasks. *Front. Comput. Neurosci.* 12, 1–10. doi:10.3389/fncom.2018.00033.

Li, X., Chen, X., Yan, Y., Wei, W., and Wang, Z. J. (2014). Classification of EEG signals using a multiple kernel learning support vector machine. *Sensors (Switzerland)* 14, 12784–12802. doi:10.3390/s140712784.

Liu, J., Li, M., Pan, Y., Wu, F. X., Chen, X., and Wang, J. (2017). Classification of Schizophrenia Based on Individual Hierarchical Brain Networks Constructed from Structural MRI Images. *IEEE Trans. Nanobioscience* 16, 600–608. doi:10.1109/TNB.2017.2751074.

Lo, A., Chernoff, H., Zheng, T., and Lo, S. H. (2015). Why significant variables aren't automatically good predictors. *Proc. Natl. Acad. Sci. U. S. A.* 112, 13892–13897. doi:10.1073/pnas.1518285112.

Loo, S. K., Lenartowicz, A., and Makeig, S. (2016). Research Review: Use of EEG biomarkers in child psychiatry research - Current state and future directions. *J. Child Psychol. Psychiatry Allied Discip.* 57, 4–17. doi:10.1111/jcpp.12435.

McGrath, J., Saha, S., Welham, J., El Saadi, O., MacCauley, C., and Chant, D. (2004). A systematic review of the incidence of schizophrenia: The distribution of rates and the influence of sex, urbanicity, migrant status and methodology. *BMC Med.* 2, 1–22. doi:10.1186/1741-7015-2-13.

Morstyn, R., Duffy, F. H., and Mccarley, R. W. (1983). Altered P300 Topography in Schizophrenia. *Arch. Gen. Psychiatry* 40, 729–734. doi:10.1001/archpsyc.1983.01790060027003.

Näätänen, R., Pakarinen, S., Rinne, T., and Takegata, R. (2004). The mismatch negativity (MMN): Towards the optimal paradigm. *Clin. Neurophysiol.* 115, 140–144. doi:10.1016/j.clinph.2003.04.001.

Neuhaus, A. H., Popescu, F. C., Bates, J. A., Goldberg, T. E., and Malhotra, A. K. (2013). Single-subject classification of schizophrenia using event-related potentials obtained during auditory and visual oddball paradigms. *Eur. Arch. Psychiatry Clin. Neurosci.* 263, 241–247. doi:10.1007/s00406-012-0326-7.

Neuhaus, A. H., Popescu, F. C., Rentzsch, J., and Gallinat, J. (2014). Critical evaluation of auditory event-related potential deficits in schizophrenia: Evidence from large-scale single-subject pattern classification. *Schizophr. Bull.* 40, 1062–1071. doi:10.1093/schbul/sbt151.

Olbrich, S., Van Dinteren, R., and Arns, M. (2016). Personalized Medicine: Review and Perspectives of Promising Baseline EEG Biomarkers in Major Depressive Disorder and Attention Deficit Hyperactivity Disorder. *Neuropsychobiology* 72, 229–240. doi:10.1159/000437435.

Park, E. J., Jin, Y. T., Kang, C. Y., Nam, J. H., Lee, Y. H., Yum, M. K., et al. (2005). Auditory and visual P300 in patients with schizophrenia and controls: Stimulus modality effect size differences. *Clin. Psychopharmacol. Neurosci.* 3, 22–32.

Sabeti, M., Katebi, S., and Boostani, R. (2009). Entropy and complexity measures for EEG signal classification of schizophrenic and control participants. *Artif. Intell. Med.* 47, 263–274. doi:10.1016/j.artmed.2009.03.003.

Santos-Mayo, L., San-Jose-Revuelta, L. M., and Arribas, J. I. (2017). A computer-aided diagnosis system with EEG based on the p3b wave during an auditory odd-ball task in schizophrenia. *IEEE Trans. Biomed. Eng.* 64, 395–407. doi:10.1109/TBME.2016.2558824.

Shim, M., Hwang, H., Kim, D., Lee, S., and Im, C. (2016). Machine-learning-based diagnosis of schizophrenia using combined sensor-level and source-level EEG features. *Schizophr. Res.* doi:10.1016/j.schres.2016.05.007.

Sonnenburg, S., Rätsch, G., Henschel, S., Widmer, C., Behr, J., Zien, A., et al. (2010). The Shogun machine learning toolbox. *J. Mach. Learn. Res.* 11, 1799–1802.

Sonnenburg, S., Rätsch, G., Schäfer, C., and Schölkop, B. (2006). Large Scale Multiple Kernel Learning. *J. Mach. Learn. Res.* 7, 1531–1565.

Speiser, J. L., Miller, M. E., Tooze, J., and Ip, E. (2019). A comparison of random forest variable selection methods for classification prediction modeling. *Expert Syst. Appl.* 134, 93–101. doi:10.1016/j.eswa.2019.05.028.

Turetsky, B. I., Dress, E. M., Braff, D. L., Calkins, M. E., Green, M. F., Greenwood, T. A., et al. (2015). The utility of P300 as a schizophrenia endophenotype and predictive biomarker: Clinical and socio-demographic modulators in COGS-2. *Schizophr. Res.* 163, 53–62. doi:10.1016/j.schres.2014.09.024.

Ulaş, A., Castellani, U., Murino, V., Bellani, M., Tansella, M., and Brambilla, P. (2012). Biomarker evaluation by multiple kernel learning for schizophrenia detection. *Proc. - 2012 2nd Int. Work. Pattern Recognit. NeuroImaging, PRNI 2012*, 89–92. doi:10.1109/PRNI.2012.12.

Varma, M., and Babu, B. R. (2009). More generality in efficient multiple kernel learning. *Proc. 26th Annu. Int. Conf. Mach. Learn. - ICML '09*, 1–8. doi:10.1145/1553374.1553510.

Wani, N., and Raza, K. (2018). "Multiple kernel-learning approach for medical image analysis," in *Soft Computing Based Medical Image Analysis* (Elsevier Inc.), 31–47. doi:10.1016/B978-0-12-813087-2.00002-6.

Woodman, G. F. (2010). A brief introduction to the use of event-related potentials (ERPs) in studies of perception and attention. *Atten. Percept. Psychophysiol.* 72, 1–29. doi:10.3758/APP.72.8.2031.A.

Yu, S., Falck, T., Daemen, A., Tranchevent, L. C., Suykens, J. A. K., De Moor, B., et al. (2010). L2-norm multiple kernel learning and its application to biomedical data fusion. *BMC Bioinformatics* 11. doi:10.1186/1471-2105-11-309.

Zhang, Y., Prasad, S., Kilicarslan, A., and Contreras-Vidal, J. L. (2017). Multiple kernel based region importance learning for neural classification of gait states from EEG signals. *Front. Neurosci.* 11, 1–11. doi:10.3389/fnins.2017.00170.

5 Funding

This work was supported by the VLIR-UOS project "A Cuban National School of Neurotechnology for Cognitive Aging"(NSNCA), Grant number CU2017TEA436A103.

6 Acknowledgments

The authors would like to thank teams of Cuban Neuroscience Center and the Department of Electronics and Informatics (ETRO) of Vrije Universiteit Brussel (VUB) for supporting this research project.

7 ANNEX

Annex 1: Feature definitions

Peak related features		
Amplitude	$A_{Peak}= max\{s(t), I_1 < t < I_2\}$, Peaks P1, P3 $A_{Peak}= min\{s(t), I_1 < t < I_2\}$, Peaks N1, N2 $[I_1, I_2]$ Detection Interval	
Latency:	$L_{Peak}=\{t \setminus s(t) = A_{Peak}\}$	
Latency/Amplitude ratio	$LAR_{Peak}= L_{Peak}/A_{Peak}$	
Absolute Amplitude	$AA_{Peak}=\lvert A_{Peak}\rvert$	
Absolute Latency/Amplitude ratio	$ALAR_{Peak}=\lvert L_{Peak}/A_{Peak}\rvert$	
Average Absolute Signal Slope	$AASS_{Peak} = \frac{1}{n}\sum_{t=I_1}^{I_2-\tau}\frac{\lvert s(t+\tau)-s(t)\rvert}{\tau}$ τ is the signal sampling period, n the number of samples of the digital signal	
Slope sign alterations	$SSA_{Peak} = \sum_{t=I_1+\tau}^{I_2-\tau}\frac{1}{2}\left\lvert\frac{s(t-\tau)-s(t)}{\lvert s(t-\tau)-s(t)\rvert}+\frac{s(t+\tau)-s(t)}{\lvert s(t+\tau)-s(t)\rvert}\right\rvert$	
Peak to Peak related features		
Peak to Peak	$PP_{Peaks} = \lvert A_{Peak} - A_{NextPeak}\rvert$	
Peak to Peak Time Window	$PPT_{Peaks} = L_{NextPeak} - L_{Peak}$	
Peak to Peak Slope	$PPS_{Peaks} = PP_{Peaks}/ PPT_{Peaks}$	
Signal related features		
Positive Area	$A_p = \sum_{t=-200}^{800}\frac{s(t)+\lvert s(t)\rvert}{2}$	
Negative Area	$A_n = \sum_{t=-200}^{800}\frac{s(t)-\lvert s(t)\rvert}{2}$	
Total Area	$A_{pn} = A_p + A_n$	
Absolute Total Area	$AA_{pn} = \lvert A_{pn}\rvert$	
Total Absolute Area	$AA_{pn} = A_p + \lvert A_n\rvert$	
Zero Crossing	$ZC_{Peaks} = \sum_{t=L_{Peak}}^{L_{NextPeak}}\delta_s$, $\delta_s = \begin{cases}1 & s(t)=0\\ 0 & s(t)\neq 0\end{cases}$	
Zero Cross density	$ZCD_{Peaks} = \frac{ZC_{Peaks}}{PPT_{Peaks}}$	
Mode frequency	$f_{mode} = f_j,\ P_j = max(P_i, 1 < i < M)$	P_j is the power spectral density of signal at a frequency bin j, M is the number of frequency bin in the spectrum
Median frequency	$\sum_{j=1}^{f_{median}} P_j = \sum_{j=f_{median}}^{M} P_j = \frac{1}{2}\sum_{j=1}^{M} P_j$	
Mean frequency	$f_{mean} = \sum_{j=1}^{M} f_j P_j \Big/ \sum_{j=1}^{M} P_j$	

8 Tables

TABLE 1. Demographic data.

	Patients	Controls	P (t-test)
Number of participants	54	54	
Male	36	36	
Age (years): mean ± std	40.5 ± 10.1	37.6 ± 14.1	0.22
Age (years): range	[22.4, 60.5]	[15.1, 64.4]	
Education (years): mean ± std	12.6 ± 1.80	14.8 ± 2.11	$4.84 \times 10{-}5$
Disease duration (years): mean ± std	14.8 ± 9.04	–	
Disease duration (years): range	[1, 40]	–	

TABLE 2. Paradigms and procedures

Auditory P300			
	tone	figure	distribution
Target	1500 Hz 70 dB	Square, side 106 pixels	10%
Distractor	500 Hz 70 dB	Circle, diameter 176 pixels	10%
Standard	1000 Hz 70 dB	Square, side 158 pixels	80%
Inter-stimulus interval was randomized between 1 and 1.5 seconds. 400 stimuli per test. 100 ms per stimuli. Total test time of 540 seconds.			
MMN			
	tone	duration	distribution
duration deviant	1000 Hz 70 dB	250 ms	5%
Frequency deviant	1500 Hz 70 dB	100 ms	5%
Standard	1000 Hz 70 dB	100 ms	90%
Inter-stimulus interval of 300 ms, 1800 tones per test. Total test time of 733 seconds			

TABLE 3. Features always selected by Boruta

PARADIGM	STIMULUS	CHANNEL	PEAK	FEATURE
P300v	Target	Pz	P2	latency
P300a	Distractor	Cz	N1	absRatio
P300a	Distractor	Fz	P2	absRatio
P300a	Distractor	Fz	P2	absAmplitude
P300a	Target	Cz	N1	absRatio
P300a	Target	Cz	N2	latency
P300a	Target	Cz	P2	latency

Table 4. Performance (%) of MKL algorithm with and without Boruta feature selection

MKL Kernels	Without FS			With FS		
	Acc	Sen	Spe	Acc	Sen	Spe
Paradigm	83	80	88	86	86	87
Channels	80	74	87	84	85	86
Type of Features	82	78	85	86	86	86

9 Figures

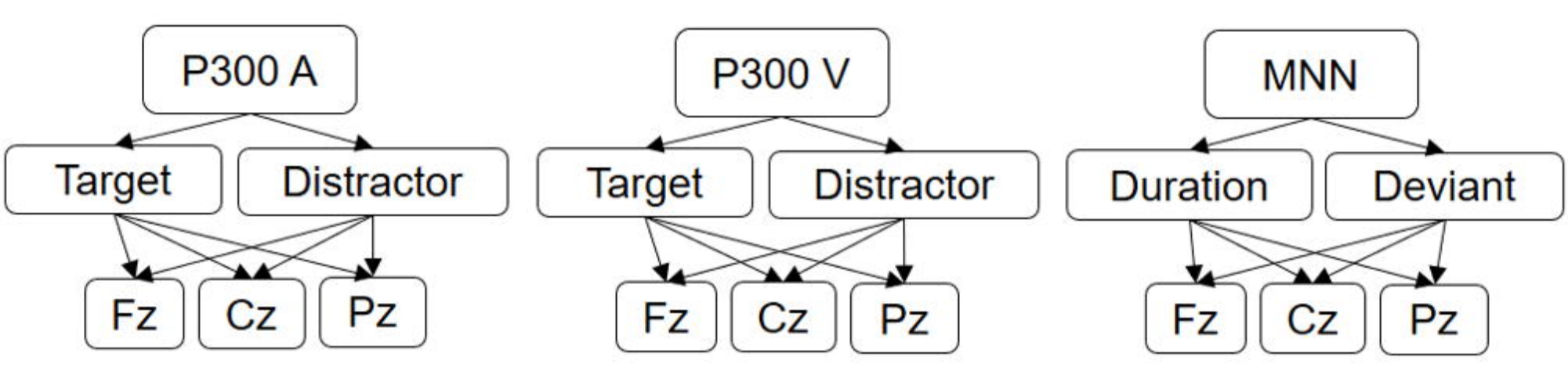
FIGURE 1. Averaged evoked potential signals used for feature extraction.

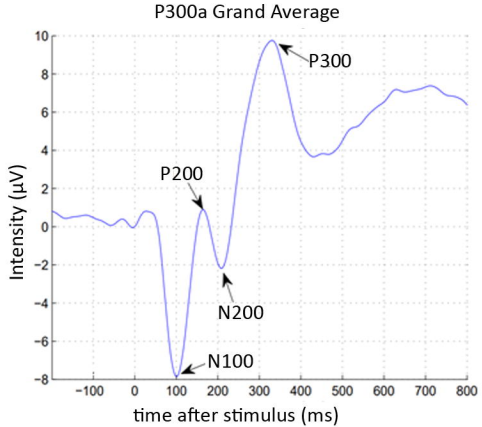
FIGURE 2. Principal components of P300 tasks (N100, P200, N200, P300) and MMN task (P200, P300).

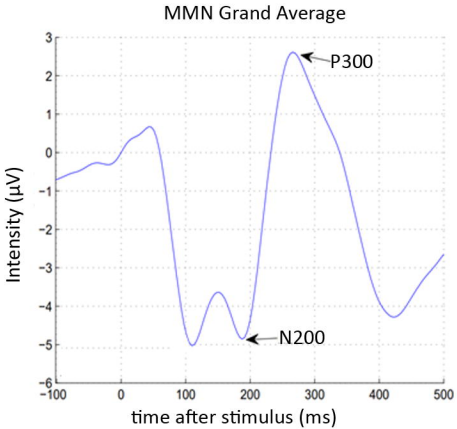
FIGURE 3. Grouping input data (726 features) in three possible kernel combinations according to the feature space approach.

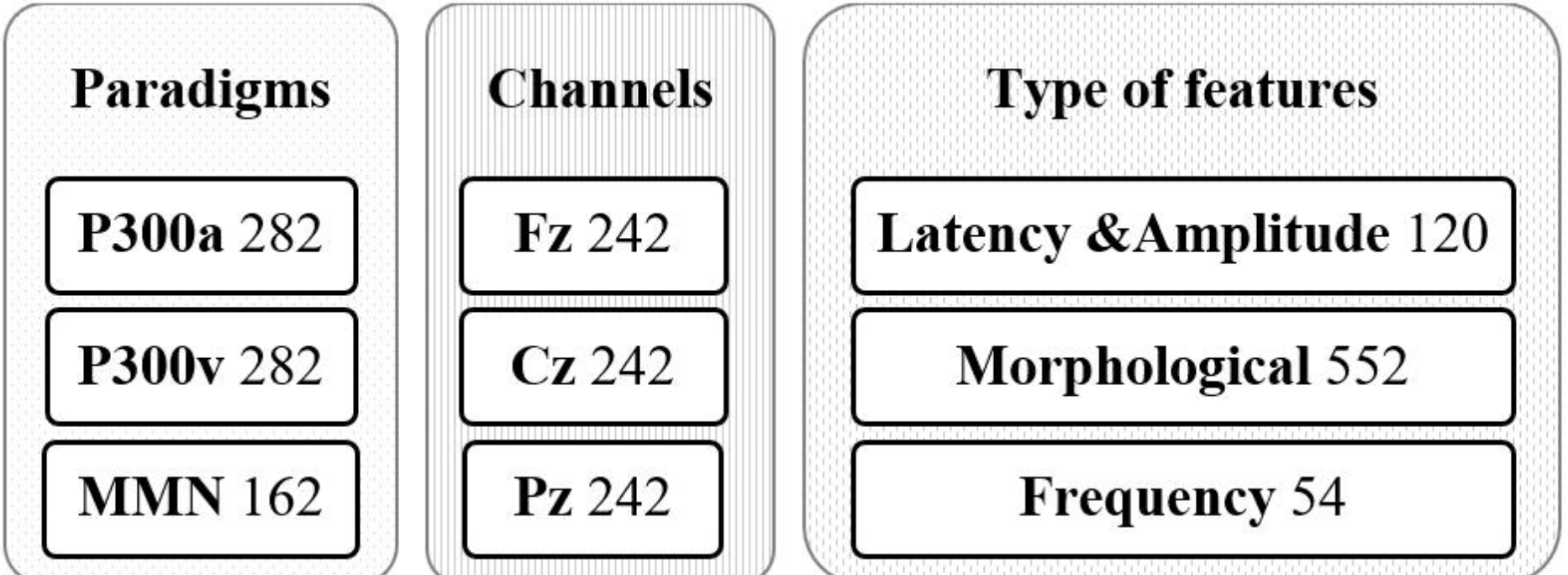
FIGURE 4. Feature selection steps applying nested cross validation.

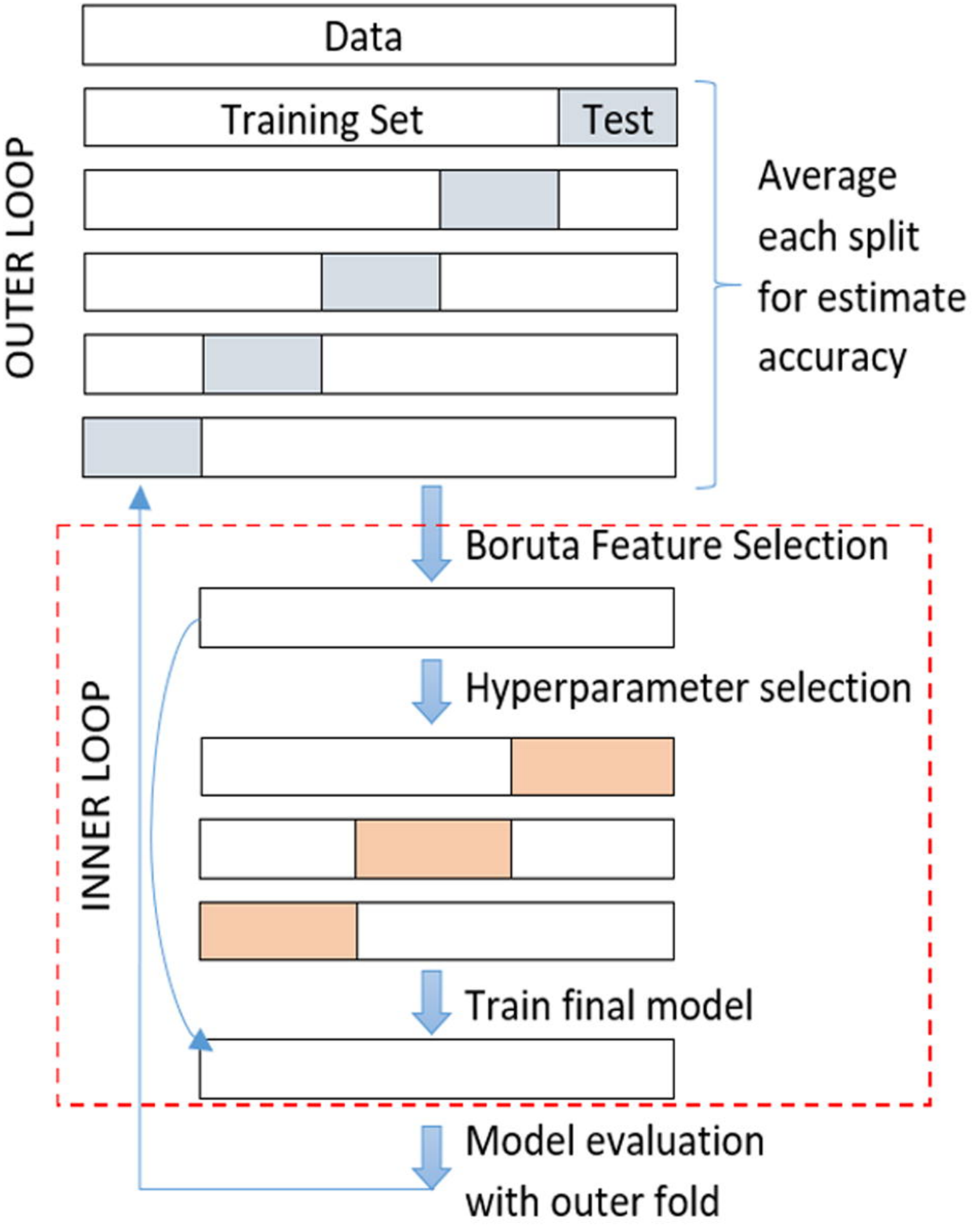
FIGURE 5. Distribution of feature selection in 10-fold-cross-validation. **(A)** Distribution per paradigm in the 10 subsets of features selected. **(B)** Frequency of selection of all the attributes that were selected in the ten Boruta applications. The bottom number means how many features were selected the number of times represented in the top number.

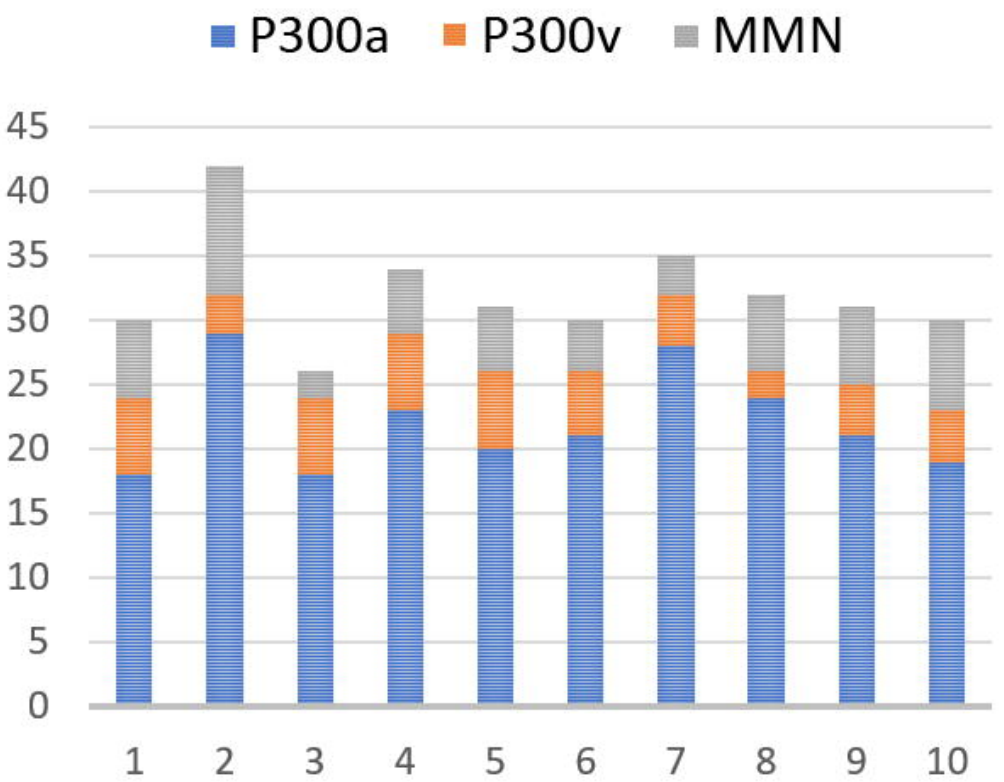

P300 A
Target
Distractor
Fz
Cz
Pz
P300 V
Target
Distractor
Fz
Cz
Pz
MNN
Duration
Deviant
Fz
Cz
Pz

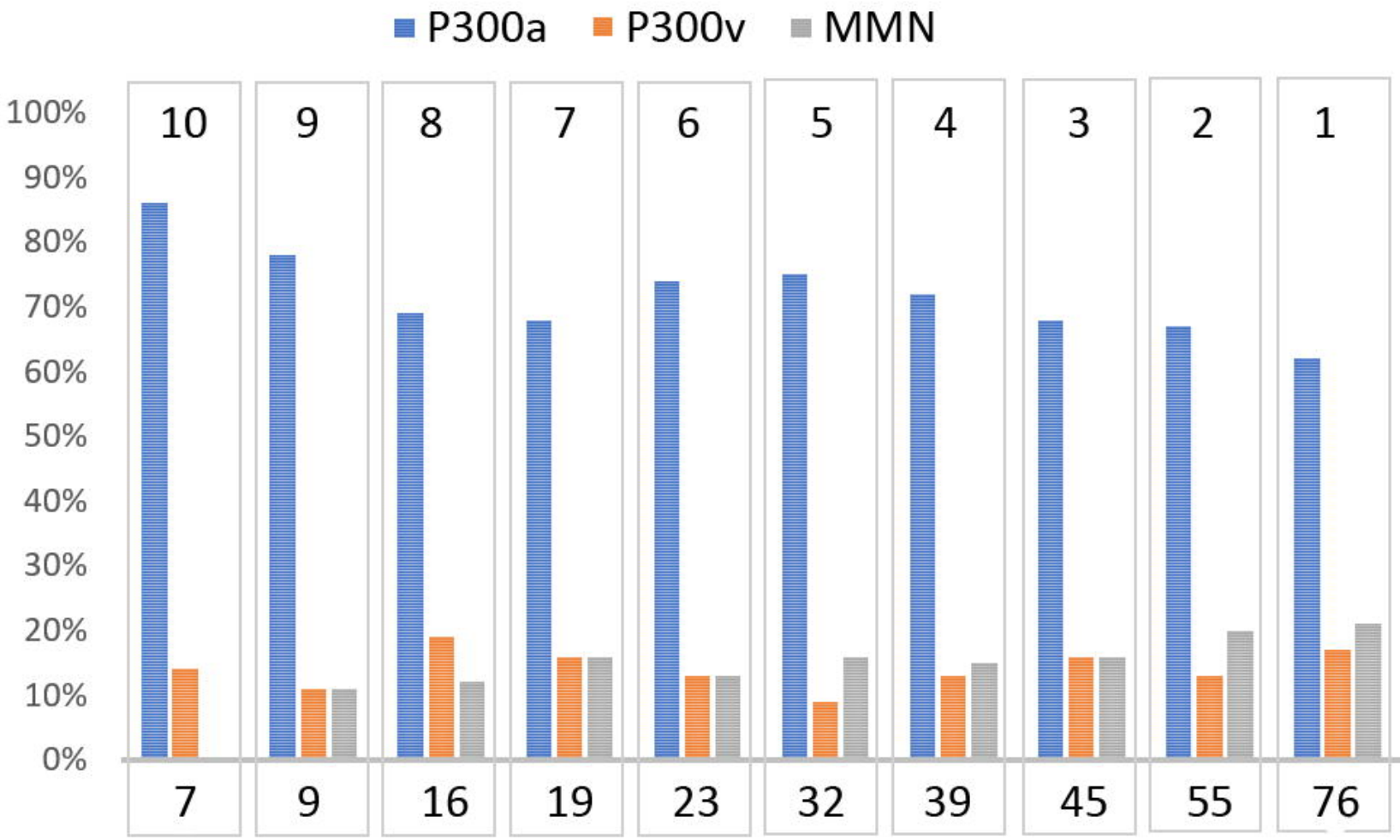

P300a Grand Average
P300
P200
N200
N100
Intensity (μV)
time after stimulus (ms)


MMN Grand Average
P300
N200
Intensity (μV)
time after stimulus (ms)


Paradigms
P300a 282
P300v 282
MMN 162
Channels
Fz 242
Cz 242
Pz 242
Type of features
Latency &Amplitude 120
Morphological 552
Frequency 54


Data
Training Set
Test
OUTER LOOP
Average each split for estimate accuracy
Boruta Feature Selection
INNER LOOP
Hyperparameter selection
Train final model
Model evaluation with outer fold


P300a
P300v
MMN
45
40
35
30
25
20
15
10
5
0
1
2
3
4
5
6
7
8
9
10


P300a
P300v
MMN
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
10
9
8
7
6
5
4
3
2
1
7
9
16
19
23
32
39
45
55
76